Интерактивное разрешение лексической и синтаксической неоднозначности в системах автоматической обработки естественного языка

Лазурский А.В., Бердичевский А.С., Крейдлин Л.Г.,
Митюшин Л.Г., Сизов В.Г.
Институт проблем передачи информации РАН
lazur@iitp.ru

Аннотация

В настоящей работе описывается метод интерактивного разрешения неоднозначности, разрабатываемый различными научными коллективами в течение последних двадцати лет [16, 18, 12]. Идея метода заключается в том, что при столкновении с неразрешимой неоднозначностью система обработки текста обращается к пользователю с уточняющим запросом. Нами созданы словари омонимов для разрешения лексической омонимии, проведены исследования, посвященные поиску возможных способов разрешения омонимии синтаксической. Полученные результаты показывают, что метод способен служить существенному повышению качества автоматического анализа и обработки текста в целом.

Abstract

The paper presents a method of interactive ambiguity resolution, which has been developed by a number of research teams for the last twenty years [16, 18, 12]. The core of the method consists in asking the user to identify a word sense, or a syntactic interpretation whenever the system lacks reliable data to make the choice automatically. We have created dictionaries of homonyms that enable the user to implement word sense disambiguation and have conducted research intended to find ways of syntactical ambiguity resolution. Our results show that the use of the method can lead to a considerable improvement of text processing quality.

0. Введение

Как известно, на данный момент ни одна система автоматического анализа и/или перевода текста не является совершенной или хотя бы близкой к таковой. Одной из основных причин неуспеха является высокий уровень неоднозначности естественного языка.

Для борьбы с неоднозначностью используется несколько основных методов: а) совершенствуются и уточняются детерминированные правила, работающие на основе лексических и грамматических данных ([1], [2], [9], [10]); б) создаются базы знаний об окружающем мире и онтологии, дающие возможность учитывать экстралингвистические данные [19]; в) разрабатываются вероятностные анализаторы, учитывающие статистические данные языка, как правило, обучающиеся в процессе работы [3], [20]. Можно заметить, что все эти методы в каком-то смысле соответствуют тем, что даны человеку: правиловый подход моделирует знание носителем правил и законов родного языка; подход, использующий базы знаний – не извлекаемые непосредственно из текста знания; статистический подход призван заменить имеющуюся у каждого человека языковую интуицию, которая позволяет справляться даже с самыми сложными случаями неоднозначности. К сожалению, все эти подходы имеют свои ограничения по эффективности и не дают результата желаемого уровня.

Таким образом, результат любого автоматического анализа/перевода должен редактироваться. Различается три типа редактирования: предредактирование, при котором текст заранее преобразуется в более понятный анализатору, в частности, снимается существенная часть неоднозначностей – примером могут послужить т. н. упрощенные, или ограниченные, языки (controlled languages); постредактирование, т. е. правка полученного на выходе текста или результата анализа и интерредактирование, т. е. редактирование непосредственно в процессе анализа, в чем и состоит идея интерактивного разрешения неоднозначности.

Помимо перечисленных выше возможностей человек располагает и другими средствами улучшения анализа текста. Например, он может уточнить смысл слов собеседника, прямо спросив его – это естественная метаязыковая операция [9], использующаяся в любых диалогах. Именно ее моделирует интерактивное разрешение, дающее машине возможность задать человеку вопрос в том случае, если она не в состоянии справиться с проблемой собственными силами.

1. Идея исследования

Идея интерактивного разрешения неоднозначности (ИРН) была выдвинута четверть века назад. По данным В. Хатчинза [16], системы МП ALPS и Weidner (США) использовали ИРН для английского языка в начале 1980-х гг. В Maruyama *et al*. [18] метод ИРН излагается применительно к японскому языку, К. Буате и Э. Бланшон в Гренобле активно развивают ИРН на материале французского, английского и других языков (МП LIDIA) [12, 13, 15]. Среди других систем обработки ЕЯ, использующих ИРН, стоит упомянуть также 1) многоязычную систему МП SYSTRAN, 2) систему ALT-J/E (Япония), 3) систему МП UMIST (Манчестер), 4) систему устного и письменного МП группы Spoken Translation (США), 5) систему многоязычного поиска и навигации в Интернете, разработанную DFKI и Университетом земли Саар (Германия).

Основными плюсами данного подхода являются простота задачи редактора (ответы на относительно несложные вопросы), универсальность (подход в целом и наша реализация в частности рассчитаны на пользователя, не имеющего специальной подготовки), высокая эффективность, широкие возможности настраивания системы для разных задач. Подход требует большой, но обозримой подготовительной работы.

Хотя интерактивное разрешение может являться лишь дополнительным средством снятия омонимии, используемым при каком-либо анализаторе, можно ожидать значительного его влияния на качество анализа/перевода. Так, например, гренобльская группа рассматривает его даже как отдельную парадигму машинного перевода – Dialogue-Based Machine Translation, наряду с Example-Based MT, Knowledge-Based MT и т. д. [19]

Особо отметим требования к пользователю (точнее, практическое отсутствие таких требований). Вопросы, которые будет задавать машина, не предполагают у пользователя ни знания лингвистических формализмов, ни специального образования, ни – в случае машинного перевода – знания выходного языка. Разумеется, обладание любым из этих качеств является плюсом и также может быть использовано, однако в обязательные требования входят лишь знание входного языка, минимальный общекультурный уровень (среднее образование) и готовность потратить некоторое время (какое именно – пользователь может решать сам).

Работа над интерактивным разрешением велась в основном именно в рамках задачи машинного перевода в системах ЭТАП и UNL и на материале этой задачи. В настоящей статье описываются

достигнутые результаты, возможность применения метода к другим задачам, а также кратко упоминаются теоретические вопросы, которые возникали в процессе работы.

2. Интерактивное разрешение лексической неоднозначности

2.1. Краткое описание

Замысел проекта состоит в том, чтобы обеспечить человека, взаимодействующего с системой автоматического анализа/перевода, простыми и ясными диагностическими описаниями неоднозначных лексических единиц, которые могли бы быть ему предъявлены на определенных стадиях обработки текста.

Основная работа, проделанная авторами, заключалась в составлении русско-английских и англо-русских «словарей омонимов» для системы ЭТАП-3. Отметим, однако, что эти или аналогичные словари могут использоваться в других задачах автоматической обработки текста, как будет показано ниже.

2.2. Принципы составления словарей

Было отобрано около 20000 русских слов, у которых леммы (или некоторые словоформы) совпадали с леммами (словоформами) других слов, и для них были написаны диагностические комментарии и примеры. Вся информация записывается в соответствующих статьях русского комбинаторного словаря. В настоящее время подобная работа осуществляется для 20000 неоднозначных английских слов.

Примеры подбираются так, чтобы максимально облегчить идентификацию значения слова. Отметим при этом, что подобрать для слова контексты, полностью исключающие возможность употребления его омонима/полисеманта, удается не всегда – «контекст определяет лексическую единицу вероятностно, а не абсолютно» [6]. В таких случаях качественные комментарии приобретают особую важность.

Комментарии могут включать: 1) аналитическое толкование значения слова или его существенный фрагмент; 2) маркер части речи, 3) простые синтаксические признаки, 4) синонимы и/или антонимы слова, – а также любые другие сведения о слове, его значении, синтактике или прагматике, которые могут оказаться полезны. В расчете на более продвинутых пользователей могут приводиться английские переводные эквиваленты. Жестких лексикографических правил

подбора комментариев и примеров нет, главная цель – обеспечить как можно более очевидное различение гипотез.

Наиболее простым случаем являются омоформы: для их различения необязательно обращаться к комментариям, достаточно предложить пользователю выбор из нескольких лемм. Рассмотрим пример:

(1) *Если увлажнить потом кожу, то эффект усиливается.*

При включенном интерактивном разрешении система задаст вопрос:

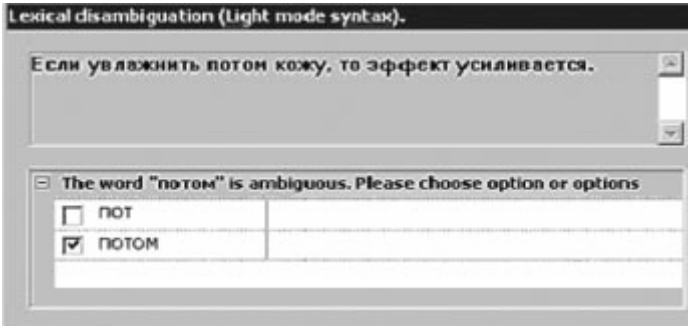

Рис 1. Фрагмент диалогового окна системы ЭТАП-3 для разрешения лексической неоднозначности: выбор лемм.

При выборе верного варианта *потом* будет выдан более или менее вероятный перевод:

(1a) *If one humidifies afterwards the skin then the effect intensifies.*

Для различения омографов, которыми являются, в частности, *по́том* и *пото́м*, можно использовать их запись с указанием места ударения. На данный момент в системе ЭТАП-3 этот метод не используется, однако реализовать его совсем несложно, т. к. в морфологическом словаре хранится информация об акцентных парадигмах слова.

Менее тривиальны случаи полисемии или омонимии[1]:

(3) *За один день Аня разучила песню из фильма «Большая перемена».*

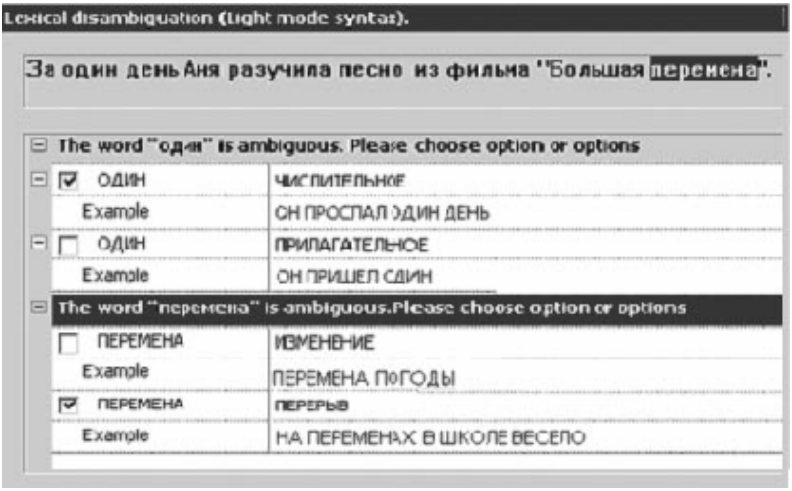

Рис 2. Фрагмент диалогового окна системы ЭТАП-3 для разрешения лексической неоднозначности: выбор полисемантов.

Выбрать верную интерпретацию для слова *один* ЭТАП сможет и сам[2], а вот для слова перемена *предпочтет* неподходящий вариант «изменение», который будет переведен как *change*. Если же пользователь укажет «перерыв», то выходом станет:

(3а) *In one day Anya learned a song from film "Large break".*[3]

Система, обогащенная интерактивным разрешением, может использовать любые знания пользователя, тем самым получая возможности, недоступные обычному анализатору. Можно позволить себе:

а) использование экстралингвистических знаний:

HASTINGS1	Noun: City in Great Britain (ГАСТИНГС)
HASTINGS2	Noun: City in New Zealand (ХЕЙСТИНГС)

Рис 3. Пример из словаря омонимов: энциклопедические сведения

В силу исторической традиции русские названия двух городов различаются. Приведенная выше запись из словаря омонимов позволит системе в нужный момент узнать у пользователя, какой именно город имеется в виду.

б) обращение к любым дополнительным признакам слова:

FAIR-HAIRED Adjective	Of usual style (СВЕТЛОВОЛОСЫЙ)
	Of high style (БЕЛОКУРЫЙ)

Рис 4. Пример из словаря омонимов: стилистические различия

в) роскошь учитывать редкие омонимы лексических единиц (скажем, английское *see* ‘епархия’) и углубляться в такие тонкости различения значений, которые чисто автоматическая система вынуждена игнорировать из-за угрозы информационного взрыва.

Важным средством экономии времени составителей словарей являются шаблоны, создаваемые для случаев регулярной омонимии, которых в языке довольно много. Если лексикограф встречает случай, который он считает регулярным, он создает для него шаблон. О шаблоне сообщается остальным разработчикам, которые уже не должны придумывать комментарий и пример для данного конкретного случая, а могут обойтись лишь подстановкой соответствующего слова в шаблон, в крайнем случае – небольшой модификацией комментария и/или примера.

EIGHTEENTH1	Adjective (ВОСЕМНАДЦАТЫЙ). Example: The **eighteenth** try
EIGHTEENTH2	Noun; 1:18 (ОДНА ВОСЕМНАДЦАТАЯ) Example: Each of eighteen guests ate an **eighteenth** of the pie

⇩

...1	Adjective (**…**) Example: The **…** try
...2	Noun; 1:**…** (ОДНА **…**) Example: Each of **...** guests ate an **…** of the pie

Рис 5. Пример из словаря омонимов: создание шаблона для омонимии порядковое прилагательное/название дроби

Модуль интерактивного разрешения неоднозначности позволяет устранить некоторые недочеты заложенных в анализатор правил, однако для этого следует тщательно продумывать взаимодействие его с другими алгоритмами: разрешения по контексту, статистического разрешения и т. п. Так, например, предложение:

(4) *Linguistics is a hard science*

получает в ЭТАПе верный (т. е. соответствующий входному тексту) перевод *Лингвистика – естественная наука.* Это достигается за счет специального правила, переводящего фразеологическое сочетание *hard science*. Но оно же действует и при переводе предложения:

(5) *Linguistics is the hardest science*,

давая в итоге *Лингвистика – естественнейшая наука,* что уже неправильно. Интерактивное разрешение здесь должно действовать в согласии с правилами перевода фразем (т.к. значения «естественный» в словарной статье слова *hard* нет). Так, должен быть задан вопрос: является ли *hard/hardest* частью идиоматического сочетания *hard science* со значением *natural science* или же это просто

свободное определение к слову наука? В случае выбора первого ответа должен быть дан перевод «естественная»/«естественнейшая», в случае выбора второго – задан вопрос, означает прилагательное *firm* (твердый) или *difficult* (трудный) (именно эти два значения записаны в словаре ЭТАПа, если словарная статья *hard* будет сложнее, то и вопросов должно быть больше).

Рассмотрим еще один пример:

(6) *You can choose either road or this picturesque footpath.*

В автоматическом режиме ЭТАП предлагает перевод *Вы можете выбрать либо дорогу, либо эту живописную тропу*. Легко убедиться в том, что этот перевод неверен: система восприняла слово *either* как часть составного союза *either…or* ‘либо… либо’. Между тем в таком случае исчисляемое существительное *road* ‘дорога’ должно было бы сопровождаться артиклем. В соответствующем правиле анализа, однако, данного требования не оказалось, что и привело к ошибке.

При включенном интерактивном модуле система предложит пользователю выбрать значение *either*, пользуясь диалоговым окном (рис. 6). Разумеется, хорошо знакомый с английским языком человек выберет вариант 1, что даст перевод *Вы можете выбрать любую дорогу или эту живописную тропу*. Добавим, что при выборе экспертом варианта 2 структура не будет построена и система перейдет в автоматический режим; наконец, выбор варианта 3 приведет к уже знакомому нам переводу. Таким образом, усилия, затраченные пользователем, вознаграждаются улучшением качества перевода.

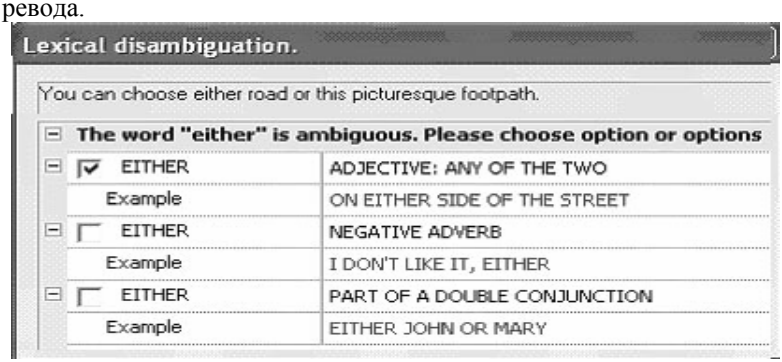

Рис 6. Фрагмент диалогового окна системы ЭТАП-3 для разрешения лексической неоднозначности

Существенно, что модуль разрешения лексической **неоднозначности** также нередко помогает справиться с **синтаксической** и **морфологической неоднозначностью**, не задавая пользователю ника-

ких вопросов о синтаксисе или морфологии: выбор правильной лексемы отсекает многие неверные варианты синтаксической структуры. Эти «побочные» эффекты возникают регулярно и расширяют возможности лексического модуля интерактивного разрешения.

2.3. Использование интерактивного разрешения лексической неоднозначности

Алгоритм анализа в ЭТАПе был модифицирован так, чтобы любой сделанный человеком выбор отсекал варианты анализа, несовместимые с ним (с возможностью возвращения к исходной ситуации, если выбор окажется тупиковым).

Было определено несколько точек процесса обработки текста, когда должно запрашиваться мнение эксперта, а именно: 1) непосредственно перед тем, как синтаксический анализатор приступает к выбору вершины дерева, 2) сразу после проверки всех гипотетических синтаксических связей, построенных правилами создания бинарных поддеревьев, 3) непосредственно перед тем, как делается выбор вариантов перевода.

Подчеркнем особо, что метод интерактивного разрешения реализуется в системе, ориентированной на получение всех возможных вариантов анализа предложения: мы не ограничиваемся одним вариантом, пусть даже наиболее вероятным.

При подобном подходе статистические методы разрешения неоднозначности отступают на второй план. Хотя система располагает целым рядом механизмов, способных подавлять маловероятные интерпретации, мы прибегаем к ним с осторожностью. Точнее говоря, система допускает два режима работы: (а) автоматический режим, при котором вероятностные соображения максимально используются для отсечения менее вероятных интерпретаций на ранних стадиях, и (б) интерактивный режим, позволяющий получить любую адекватную интерпретацию. В этом режиме роль статистических соображений не сводится к нулю, но становится менее приоритетной.

Все созданные словари существуют в электронном виде и полностью готовы к использованию.

3. Интерактивное разрешение синтаксической неоднозначности

Синтаксическая неоднозначность представляет собой гораздо более сложный случай для интерактивного разрешения, чем лекси-

ческая. Как показывает опыт, рядовой пользователь легко различает лексические значения, но не очень готов к ответу на синтаксические вопросы. Таким образом, основная задача создания модуля синтаксического разрешения заключается в визуализации информации – наглядном представлении различных синтаксических конструкций [7]. С другой стороны, потенциальная мощность синтаксического разрешения неоднозначности в целом выше мощности лексического. Последнее неизбежно реализуется в частных правилах: конкретно, в использовании словарных данных.

В принципе можно ожидать, что для синтаксической омонимии реально создать более глобальные правила, которые будут справляться со многими ее случаями, т.е. быть более универсальными. Однако для этого надо построить соответствующее исчисление возможных случаев синтаксической неоднозначности, что нам пока что не удалось. По всей вероятности, разработать абсолютно универсальные правила наглядного представления произвольных типов синтаксической неоднозначности невозможно. Мы постарались изучить, какого рода запросы наиболее понятны пользователю и определить некоторые частотные типы синтаксической неоднозначности, для которых стоит составлять специальные правила наглядного представления.

3.1. Существующие алгоритмы интерактивного разрешения синтаксической неоднозначности

На данный момент ЭТАП-3 предоставляет возможность интерактивного синтаксического разрешения специалистам, хорошо знакомым с системой. Ведется диалог, позволяющий пользователю выбирать между синтаксическими гипотезами, имеющими вид бинарных поддеревьев. Такой метод особенно эффективен, если разрешение лексической и синтаксической неоднозначности производится для предложения одновременно.

Рассмотрим пример:

(7) *Дом портит засоренный жильцами мусоропровод.*

Анализ этого предложения в чисто автоматическом режиме (при условии выдачи всех вариантов перевода) занимает 1 минуту 25 секунд, выходом являются 16 различных вариантов, из которых первым (т. е. тем, который система считает наиболее вероятным) идет

(7а) *The house spoils a refuse chute, littered by the tenants.*

Во-первых, из-за инвертированного порядка слов неверно идентифицированы синтаксические отношения. Переводы с верной структурой (в которых подлежащим является мусоропровод) появ-

ляются лишь начиная с девятого. Во-вторых, не очень удачно выбран перевод для *засорять*, *litter* – скорее «сорить». Другой вариант перевода, имеющийся в словаре ЭТАПа – *choke up* – подходит гораздо лучше. Еще одной «точкой неоднозначности» является перевод синтаксического отношения *засоренный → жильцами* (жильцы – те, кто засорил, или жильцы – те, чем засорено). Три бинарных точки неоднозначности плюс вариативность перевода настоящего времени глагола *портит* – *spoils* или *is spoiling* - как раз и дают 16 вариантов перевода.

При подключении модуля интерактивного разрешения появляется следующее окно:

Рис 7. Диалоговое окно системы ЭТАП-3 для разрешения синтаксической неоднозначности: формализм поверхностно-синтаксических отношений

Как видно, вопросы формулируются в терминах грамматики зависимостей и формализма поверхностно-синтаксических отношений, принятого в модели «Смысл⇔Текст». Первый и третий вопрос являются, по сути, одним и тем же: какое из слов *дом* и *мусоропровод* – подлежащее, а которое – прямое дополнение, достаточно ответить лишь на один, ответ на второй в силу неповторимости соответствующих синтаксических отношений будет получен компьютером автоматически.. Второй вопрос уточняет роль, которую жильцы играют в процессе засорения.

Затем появляется окно лексического разрешения:

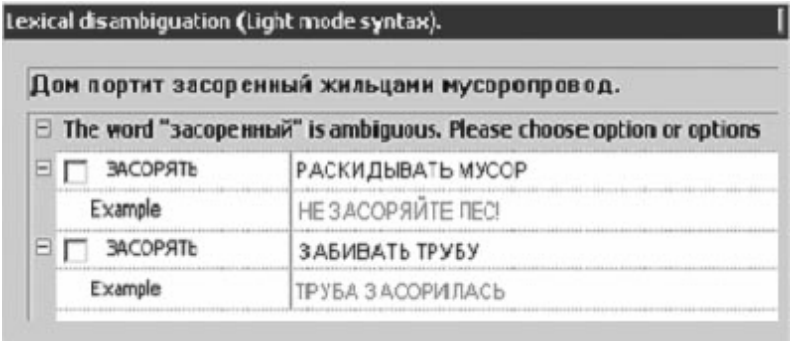

Рис 8. Фрагмент диалогового окна системы ЭТАП-3 для разрешения лексической неоднозначности

Выбор верного варианта уменьшает количество возможных интерпретаций (после двух предыдущих вопросов их осталось четыре) вдвое, таким образом, их остается всего две, различающиеся лишь употреблением Present Simple или Present Continuous. Первым выдается вполне приличный вариант:

(7б) *It is a refuse chute, choked up by the tenants, that spoils the house.*

Как видно, сочетание модулей лексического и синтаксического разрешения значительно повышает качество перевода и уменьшает время работы – для ответа на вопросы опытному пользователю требуется всего лишь несколько секунд.

3.2. Некоторые специальные правила для наглядного представления синтаксической информации

Рассмотрим частый случай неоднозначности в английском языке: существительное с несколькими предшествующими ему несогласованными определениями:

(8) *Fat soup admirer.*

ЭТАП предлагает десять вариантов анализа и перевода этой вполне безобидной фразы:

Откормите поклонника супа;
Поклонник жирного супа;
Жирный поклонник супа;
Поклонник супа жира;
Поклонник супа жира.

Остальные пять вариантов отличаются тем, что в них употребляется слово *поклонница*. Четвертый и пятый варианты различаются древесной структурой:

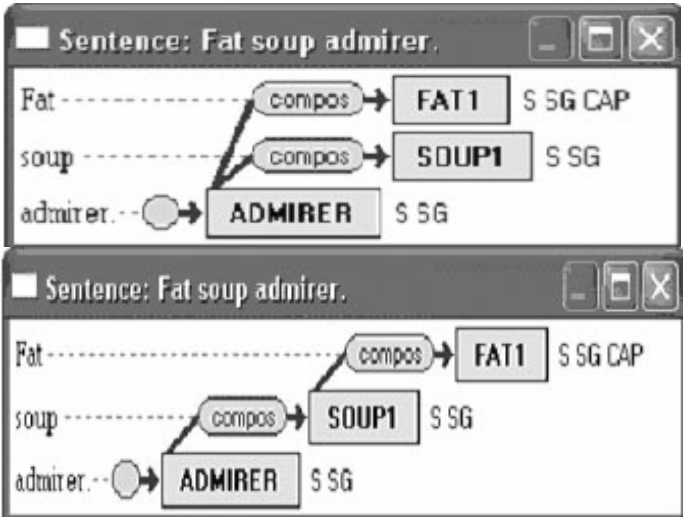

Рис 9. Две из синтаксических структур для предложения *Fat soup admirer.*

Как видно, интерактивное разрешение тоже несет в себе определенную угрозу информационного взрыва, так как всегда есть опасность засыпать пользователя вопросами. В данном случае вопросы для различения четвертого и пятого варианта абсолютно не нужны, так как на перевод различия структур не влияют, но если анализатор используется не для перевода а, например, для разметки корпуса, то различия важно учесть. К счастью, все варианты, кроме второго и третьего, могут быть отвергнуты с помощью модуля разрешения лексической неоднозначности (вообще говоря, различение английских прилагательных и существительных, необходимое для отсечения четвертого и пятого вариантов – задача непростая и решаемая в известной степени условно, но в данном случае проблемы не возникает). Вопрос для выбора между вторым и третьим вариантами можно задать, выделяя минимальные поддеревья:

The given sentence is ambiguous. What does the word *fat* refer to?

Fat soup

Fat admirer

Соответствующее правило может выглядеть как: (I) «Если деревья различаются тем, что слово X имеет в них разных хозяев, то для

каждого дерева написать рядом слово X и его хозяина. Задать вопрос о том, к чему относится слово X».

Можно использовать и другое правило: (II) «Если деревья различаются тем, что слово X имеет в них разных хозяев, то для каждого дерева объединить слово X и его хозяина в линейной структуре предложения скобочками. Отдельно выделить скобочками хозяина X, если в линейной структуре хозяин расположен не рядом с X, выделить скобочками составляющую, в которую он входит». Это правило представляется более сложным, а результат его работы – менее наглядным:

The given sentence is ambiguous. How should it be interpreted?

(Fat (soup)) admirer

(Fat (soup admirer))

однако для некоторых предложений оно может оказаться более подходящим, чем правило I.

Рассмотрим предложение:

(9) *He studies buzzes and whistles*.

При анализе предложение законно получает две возможные структуры:

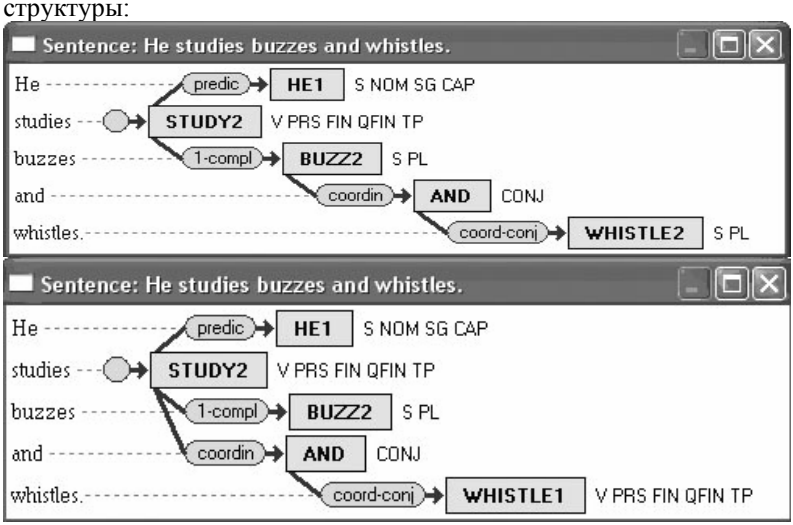

Рис 10. Две синтаксические структуры для предложения *He studies buzzes and whistles*.

и, соответственно, два перевода: *Он изучает жужжание и свист; Он изучает жужжание и свистит.* Неоднозначность заключается в сочинительной связи – с чем сочиняется *whistles*: с *buzzes* или со *studies*? Результатом работы несложного правила «Если структуры

различаются тем, что слово X в них соединено сочинительной связью с разными словами, то для каждой структуры объединить X с его сочинительным партнером при помощи соответствующего сочинительного союза» станет вопрос:

The given sentence is ambiguous. How should it be interpreted?

buzzes and whistles

studies and whistles

Дальнейшая разработка модуля интерактивного разрешения синтаксической неоднозначности должна заключаться именно в совершенствовании подобных правил, с одной стороны – в поиске возможностей их объединения в более глобальные правила, с другой – напротив, в их детализации и учете возможных исключений для каждого правила.

Параллельно возможно использовать также реализованную в системе ЭТАП-3 систему русского перефразирования, стремясь получить несколько выражений, синонимичных исходному, однако менее неоднозначных.

3.3. Достоинства и недостатки интерактивного разрешения синтаксической неоднозначности по сравнению с лексической

Можно заметить, что синтаксическое разрешение менее универсально в смысле практической применимости: если словари для лексической омонимии можно подключить к любому автоматическому обработчику, то правила для синтаксической зависят от анализатора и типа используемого формализма. Так, например, все приведенные выше правила сформулированы в терминах грамматики зависимостей (хотя в принципе могут быть приспособлены и к другому формализму).

Кроме того, синтаксические представления менее понятны пользователю. Проведенные нами испытания показывают, что средний пользователь хотя и нередко, но не всегда понимает вопросы с переформулировкой предложения, выделение составляющих скобками, использование фрагментов дерева зависимостей и т. п.

С другой стороны, в случае «чистой» синтаксической неоднозначности другие средства могут оказаться бессильны. Выбор правильной древесной структуры может оказаться очень существенным как для машинного перевода, так и для менее прикладных задач (к примеру, для синтаксической разметки корпуса русского языка использовался синтаксический анализатор ЭТАПа, причем велось постоянное интерредактирование). Соответственно, при реа-

лизации определенных задач интерактивное синтаксическое разрешение должно применяться, несмотря на его сложность.

4. Настройка и самонастройка системы

Очевидно, что система интерактивного разрешения может использоваться при решении разных задач: автоматический перевод, разметка корпусов, формулирование поискового запроса, теоретические исследования и т. д. Соответственно, уровень подготовки пользователей, их намерения и время, которое они будут готовы посвятить работе, также будут различаться. На наш взгляд, интерфейс системы должен предоставлять пользователю следующие возможности настраивания:

- Степень участия пользователя (автоматический режим vs. интерактивный (см. выше) – разумеется, режимов необязательно должно быть именно два, скорее, должна предлагаться некоторая шкала). От этого зависит, на какой стадии система начнет обращаться к пользователю, насколько позволит себе «эксплуатировать» его знания и т. п.
- Задача пользователя: перевод (в случае перевода – с указанием языков), разметка, другое.
- Уровень подготовки пользователя: знание морфологических терминов[4], знание лингвистических формализмов, знание выходного языка (в случае перевода).
- Степень игнорирования несущественной и/или трудноустранимой омонимии.

Важным элементом интерфейса мы также считаем «обойму» вопросов: пользователю задается не один вопрос, а сразу несколько, с тем, чтобы он мог выбрать тот, на который ему легче всего ответить, – а ответ на один вопрос, как можно было видеть выше, часто снимает другие (такое предложение первым высказывал H. Blanchon).

Стоит отметить, что в процессе диалога с человеком система получает от него большое количество информации. Разумеется, возникает желание эту информацию использовать не только для «сиюминутного» разрешения», но и запоминать ее и накапливать для последующего автоматического использования.

Так, например, если пользователь указал для себя высокий уровень подготовки, система может использовать его для пополнения своих словарей: в случае отсутствия в словаре нужного значения давать пользователю возможность ответить **Другое** и вводить соответствующее значение, добавляя к нему толкование и пример (если

пользователь считает, что контекст, в котором встретилось неоднозначное слово, является хорошим характеризующим примером, можно просто запоминать его).

5. Некоторые применения интерактивного разрешения неоднозначности в интернет-сервисах

5.1. Система UNL

Одной из перспективных разработок, направленных на развитие интернет-технологий и использующих методы интерактивного разрешения, является международный проект UNL. Проект ставит перед собой цель частично преодолеть языковой барьер, разделяющий пользователей Интернета. Суть проекта заключается в том, что предлагается универсальный язык-посредник, достаточно мощный для того, чтобы на нем можно было выразить всю важнейшую информацию, которую передают тексты на естественных языках. Этот язык – Универсальный Сетевой Язык (Universal Networking Language, или UNL) предложил Х. Учида (Университет ООН). Для каждого естественного языка предлагается разработать две системы: «деконвертор», который переводил бы тексты с языка UNL на данный язык, и «энконвертор», который преобразовывал бы тексты на данном языке в выражения языка UNL. Следует подчеркнуть, что порождение текста на языке UNL не будет полностью автоматическим. Эта процедура планируется как **диалог** между компьютером и человеком (редактором).

Таким образом, данный проект принципиально отличается от традиционного машинного перевода. Прежде всего, входом для порождения текстов на разных естественных языках, служит структура UNL, качество которой не зависит от несовершенства процедур анализа текстов. В процессе интерактивного построения UNL структуры редактор будет просматривать результаты работы автоматического энконвертора, исправлять ошибки и разрешать оставшуюся многозначность. Затем редактор может запустить деконвертор и перевести отредактированное им UNL выражение на свой родной язык, чтобы проверить результаты своей работы и при необходимости внести в это выражение дополнительные изменения.

UNL – это компьютерный язык, разработанный для представления информации в таком виде, который позволял бы порождать тексты, содержащие эту информацию, на самых разнообразных языках. Для деконвертации с UNL на русский и энконвертации с русского на UNL используются соответствующие модули системы ЭТАП-3.

При установлении соответствий между русскими словами и минимальными лексическими единицами UNL (т. н. универсальными словами) особо важно максимально избавиться от лексической неоднозначности, с тем чтобы получить как можно более точное UNL-представление русского предложения. В качестве одного из основных механизмов для этого используются несколько модифицированные словари омонимов.

5.2. Формулирование поискового запроса

Как уже упоминалось, использование интерактивной неоднозначности (главным образом лексической) может оказаться полезным для повышения точности интернет-поиска. В самом деле, в набор найденных ссылок за счет неразличения омонимов часто попадают совершенно нерелевантные: к примеру, человек, желающий получить информацию о количестве узких оборонительных рвов вокруг Москвы в годы Великой Отечественной войны и задающий Яндексу запрос типа: *«количество траншей» Москва война*, получит среди десяти первых результатов лишь один релевантный, остальные девять будут содержать подробные описания облигационных займов.

Очевидно, используя те же словари лексической неоднозначностей, легко можно попросить пользователя уточнить запрос и тем самым выяснить, имеется в виду *траншея* или *транш*, человек Владимир Ковров или поезд Владимир–Ковров и т. п.

Разумеется, способ будет эффективен лишь в том случае, если есть возможность снимать неоднозначность и в результатах поиска или вообще во всем поисковом пространстве. Это, впрочем, вполне реально: ведь можно воспользоваться различными механизмами уменьшения неоднозначности: вероятностными алгоритмами, поиском в текстах наиболее соответствующей запросу тематики (при условии предварительной рубрикации), простым разрешением по ближайшему контексту и любыми другими методами, разбор которых сейчас не входит в нашу задачу: важно, что они существуют, а большой объем и связность текста позволяют их применять. Применить же соответствующие алгоритмы – как статистические, так и правиловые – к короткому и «бессвязному» запросу представляется малоперспективным.

6. Теоретические вопросы: внутриязыковая и переводная неоднозначность

Важным аспектом нашего подхода является то, что мы отдельно рассматриваем внутреннюю неоднозначность входного языка и неоднозначность, возникающую при переводе. Это различие особенно важно проводить в случае многоязычной системы, каковой является, например, система UNL (см. раздел 5.).

Действительно, если одни случаи неоднозначности не зависят от того, на какой язык переводится текст (и, шире, не зависят от конкретной задачи обработки ЕЯ – как, например, словосочетание *простой солдат*), то неоднозначности другого типа возникают только при переводе на определенный язык. Например, при переводе с русского языка на английский слову *свеча* (в прямом значении) соответствует слово *candle*, в то время как при переводе на французский следует учесть материал, из которого сделана свеча, и соответственно перевести это слово как *la chandelle* (сальная свеча), *la bougie* (стеариновая) или *le cierge* (восковая).

Поскольку эти типы неоднозначности имеют разный характер, они разрешаются на разных стадиях обработки предложения: внутриязыковая неоднозначность - во время анализа предложения, а переводная – на этапе собственно перевода[5]. Если бы это различие не учитывалось и оба типа рассматривались одновременно на стадии анализа, то пришлось бы нагрузить описание входного языка информацией обо всех неоднозначностях всех выходных языков, что трудоемко и крайне неестественно. С другой стороны, если бы разрешение внутриязыковой неоднозначности было отложено до стадии перевода, мы лишились бы возможности раннего отсечения неправильных интерпретаций. Насколько нам известно, ЭТАП-3 – единственная система, проводящая четкое различие между указанными типами неоднозначности.

7. Результаты и выводы

Из вышесказанного видно, что метод интерактивного разрешения может оказываться эффективным там, где полностью автоматические методы бессильны, и существенно повышать качество автоматической обработки текста. Метод рассчитан на выбор **наилучшего** варианта из нескольких возможных, но может оказаться полезным и при отсечении заведомо неправильных вариантов.

Стоит, однако, отметить, что интерактивное разрешение не может являться основным средством анализа и предназначено лишь для выполнения функций надстройки дизамбигуатора. Таким образом, если качество основного анализатора заведомо низкое, существенного повышения уровня за счет интерактивного разрешения не произойдет.

На данный момент более эффективным и простым представляется разрешение лексической неоднозначности, однако разрешение синтаксической является также возможным и необходимым.

В процессе работы приходится как решать практические задачи, так и задумываться на теоретическими вопросами – например, такими, как составление лексикографических правил заполнения словарей омонимов, различение внутренней и переводной неоднозначности, несовпадения границ омонимов в разных языках.

В завершение хотелось бы подчеркнуть, что метод интерактивного разрешения неоднозначности языковых единиц, несмотря на некоторую прямолинейность и тот факт, что он заметно уступает другим методам в тонкости и изощренности, оказывается весьма перспективным и заслуживает внедрения в самых разных областях автоматической обработки текстов.

Список литературы

1. Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Н. В. Перцов, В. З. Санников, Л. Л. Цинман. *Лингвистическое обеспечение системы ЭТАП-2*. Москва, Наука (1989). 295 стр.
2. Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, В. З. Санников, Л. Л. Цинман. *Лингвистический процессор для сложных информационных систем*. Москва, Наука (1992), 256 стр.
3. И. М. Богуславский, Л. Л. Иомдин, В. Г. Сизов, И. С. Чардин. *Использование размеченного корпуса текстов при автоматическом синтаксическом анализе*. // Труды международной конференции «Когнитивное моделирование в лингвистике-2003». Варна (2003), стр. 39-48.
4. И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, А. С. Бердичевский, Л. Г. Крейдлин, В. Г. Сизов. *Интерактивное разрешение неоднозначности различных типов в машинном переводе* // Труды международной конференции Диалог'2005. Москва, 2005.
5. И. М. Богуславский, Л. Л. Иомдин, Л. Г. Крейдлин, Н. Е. Фрид, И. Л. Сагалова, В. Г. Сизов. *Модуль универсального сетевого языка*

(UNL) в составе системы ЭТАП-3 // Труды международной конференции Диалог'2000. Москва, 2000.
6. Лайонз Дж. *Введение в теоретическую лингвистику*. М., 1977.
7. Мальковский, М. Г. Осин, А. И. *Визуализация смысла сложных для восприятия фрагментов текста.* Труды международной конференции Диалог'2001. Москва, 2001.
8. И. А. Мельчук. *Опыт теории лингвистических моделей класса «Смысл ⇔ Текст»* Москва, Наука (1974).
9. Р. О. Якобсон. *Шифтеры, глагольные категории и русский глагол* // Принципы типологического анализа языков различного строя. М., 1972.
10. Apresian, Ju. D., Boguslavsky, I. M., Iomdin, L. L., Lazursky, A. V., Sannikov, V. Z, Sizov, V. G., Tsinman, L. L. *ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT*. // MTT 2003, First International Conference on Meaning – Text Theory. Paris, École Normale Supérieure, Paris, June 16–18, 2003, pp. 279-288.
11. Apresjan, Ju. D., Boguslavskij, I. M., Iomdin, L. L., Lazurskij, A. V., Sannikov, V. Z. and Tsinman, L.L. *Système de traduction automatique {ETAP}*. La Traductique. P.Bouillon and A. Clas (eds). Montréal, Les Presses de l'Université de Montréal. (1993).
12. Blanchon, H. *An Interactive Disambiguation Module for English Natural Language Utterances*. // Proceedings of NLPRS'95. (Seoul, Dec 4-7, 1995), vol. 2/2: 550-555.
13. Blanchon, H. & Fais, L. (1997). *Asking Users About What They Mean: Two Experiments & Results*. Proc. HCI'97. San Francisco, California. August 24-29, 1997. vol. 2/2: pp. 609-912.
14. Boguslavsky, Igor M., Iomdin, Leonid L., Lazursky, Alexander V., Mityushin, Leonid G., Sizov, Victor G., Kreydlin, Leonid G., Berdichevsky, Alexander S. *Interactive Resolution of Intrinsic and Translational Ambiguity in a Machine Translation System*. // CICLing 2005. Lecture notes in computer science. A.Gelbukh (ed.), Springer-Verlag Berlin Heidelberg 2005, pp. 383 - 394.
15. Boitet, C., Blanchon, H. *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup*. // Machine Translation, 9/2 (1994), 99-132.
16. Hutchins W. *Machine translation: past, present, future*. Ellis Horwood, Chichester (1986).
17. Iomdin, L. L, Sizov, V. G, Tsinman, L.L. *Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique*. META, 47. (3). (2002) 351-358
18. Maruyama H., Watanabe, H., and Ogino, S. *An interactive Japanese parser for machine translation*. // Karlgren, H., ed. Proceedings of the

13th International Conference on Computational Linguistics, v. 2. Helsinki. (1990). 257-62
19. Nirenburg S., Raskin, V. *Ontological Semantics.* http://crl.nmsu.edu/Staff.pages/Technical/sergei/book/ // Computing Research Laboratory.
20. Tufis D., Ion, R., Ide, N. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. // Proceedings of the 20th International Conference on Computational Linguistics, Geneva, August 23–27, 2004, pp. 1312-1318.
21. *Universal Networking Language (UNL) Specifications. Version 2005.* http://www.undl.org/unlsys/unl/unl2005/ // Universal Networking Digital Language Foundation.

---

[1] Полисемичные и омонимичные лексемы по-разному оформляются в словарях, однако никакого влияния на ведение диалога это не оказывает.

[2] Хотя даже в этом случае нельзя считать применение интерактивного разрешения неоправданным. Как будет показано ниже, человек нередко затрачивает на выбор ответа меньше времени, чем потратила бы система на автоматический поиск верной интерпретации. Кроме того, ответ освобождает ее от необходимости тратить время на анализ и перевод неверных или менее подходящих интерпретаций: ЭТАП выдает списком все переводы, которые считает возможными, за счет интерактивного разрешения этот список нередко сокращается во много раз, в идеале – до одного-единственного варианта.

[3] *Large* – не очень удачно, но это уже недостаток данных о сочетаемости слова *break*.

[4] В случае, если пользователь хорошо разбирается в грамматических категориях, можно будет разрешать морфологическую омонимию вопросами «в лоб» типа «какой это падеж?». Подобный диалог (с комментариями для неуверенных пользователей) остроумно реализован, например, в Конкордансере Сидорова.

[5] В системе ЭТАП-3 результат работы блока анализа – нормализованная синтаксическая структура – поступает на вход блока перевода, а затем итог работы данного блока передается блоку синтеза выходного языка. Разграничение между этапами весьма жесткое.